中文评论产品特征与观点抽取方法研究*

孟　园　王洪伟
(同济大学经济与管理学院　上海 210000)

摘要: 【目的】针对中文在线评论产品特征与观点抽取问题, 提出一种基于置信度排序模型的抽取方法。【方法】在改进 HITS 算法基础上, 综合考虑候选特征观点词的关联关系和语义关系构建置信度排序模型, 提取并过滤特征观点词。【结果】和基准模型相比, 本文方法对中文语料的产品特征和观点抽取能达到较高准确率和召回率。【局限】仅针对产品显性特征抽取, 没有考虑隐性特征的识别与抽取。【结论】利用特征词和观点词的双向增强关系和语义关系, 可以有效抽取产品特征观点; 情感极性过滤对提升观点词抽取准确率有较大作用。

关键词: 置信度排序　HITS　关联关系　语义关系　双向增强关系　特征观点抽取
分类号: G350

1　引　言

互联网环境日益成熟, 越来越多的消费者倾向于通过电商网站进行购物并点评, 由此产生了数据量庞大的在线评论。研究表明, 从大量的点评信息中提取针对产品特征的评价观点尤其重要, 它不仅便于消费者迅速了解产品各方面性能, 判断产品质量; 更为企业提供了产品设计的依据和其他企业的竞争情报, 促进企业竞争力的提升[1]。可见, 抽取评论中的产品特征及其评价观点具有重要的商业价值, 因而成为情感分析领域关键的研究任务之一。

常见的用户评论中(如外观非常*漂亮*, 外观很*不错*。速度*不错*。), 观点词通常出现在特征词的附近, 用来描述或修饰产品特征, 两者具有较强的关联性。假设名词为候选特征词, 形容词为候选观点词, 不难发现, 一个可以被越多不同的观点词修饰的名词, 越有可能是特征词(如"外观")。相似地, 一个可以修饰越多不同特征词的形容词, 越有可能是观点词(如"不错")。这种候选特征词和候选观点词之间相互影响的关系, 称为双向增强关系。利用这一关系, 文献[2]和文献[3]引入排序算法, 计算候选词的置信度, 最后抽取置信度达到阈值的候选词作为正确的特征词或观点词, 取得了一定效果。然而, 现有的相关研究中, 常常忽视了词语的语义关系和关联关系对于抽取结果的影响作用。例如, 如果确定"外观"是正确的特征词, 那么候选集中与"外观"语义相近的其他词语"外形"、"外表"等, 也更可能成为特征词。不仅如此, 经常一起搭配出现的名词和形容词往往更有可能成为正确的特征观点词(如"价格"和"贵")。

为此, 本文基于 HITS 排序算法, 综合考虑候选特征观点词对的关联关系, 以及特征词或观点词间的语义关系, 构建置信度排序模型抽取产品特征及观点。同时, 还采取不同的策略对特征词和观点词进行过滤, 取得了较好的实验结果。

2　相关文献综述

目前产品特征及观点的抽取方法主要分为监督学习方法和非监督学习方法。

(1) 监督学习方法

Jin 等[4]采用 HMMs 模型识别特征词、观点词及观点极性。Li 等[5]整合了 Skip-CRF 和 Tree-CRF 提取评价对象。Wu 等[6]采用 SVM 分类器, 根据短语依存关

通讯作者: 孟园, ORCID: 0000-0002-6595-8370, E-mail: nancymeng5544@163.com。

*本文系国家自然科学基金项目"中文语境下基于模糊本体的用户在线评论的情感分析"(项目编号:70971099)和国家自然科学基金项目"在线评论对商家业绩的影响研究:情感分析的视角"(项目编号:71371144)的研究成果之一。

系发现评价对象和评价词语之间的关系。由于监督学习方法依赖于大量的人工标注工作的准确性，且领域独立性较差，在实际领域中的应用仍存在诸多限制。近年来，学者们积极探索各种非监督学习方法抽取特征观点词。

(2) 非监督学习方法

主流方法包括主题建模方法和语料统计方法。

Titov 等[7]提出多粒度主题模型，应用于文档中连续的数条句子，得到按主题自动聚类的特征词和观点词及其多项分布。Zhao 等[8]提出 MaxEnt-LDA 为产品特征及观点词联合建模，并使用句法特征辅助两者分离。主题模型可以用于多种信息建模，扩展性强，但在实际中，实验结果并不稳定，并且很难发现在局部文档中频繁出现的特征词。因此，一些学者倾向于语料统计方法获取特征观点词。Hu 等[9]利用关联规则算法，抽取名词中的频繁项集作为候选特征词，并利用最近邻原则抽取距离频繁名词或名词短语最近的形容词作为观点词。这种方法将名词作为候选特征词，容易产生大量无关特征词。后续，Aravindan 等[10]采用近邻规则(Compactness Rule)和独立支持度(P-Support)规则进行过滤改进。Qiu 等[11]、Hai 等[12]基于双向传播算法，利用特征观点词的依存关系或关联关系，通过特征词抽取观点词、观点词抽取特征词的双向传播方式，迭代抽取更多新的特征词和观点词，直至结束。实验结果取得了较高的召回率，但随着迭代的深入产生了较多噪音词，准确率不高。还有一些学者基于排序算法，利用特征指示词和特征词之间的双向增强关系迭代计算，最后抽取出置信度高的候选特征词作为正确的产品特征和观点词，取得了较好的效果，如 Zhang 等[2]、郗亚辉[3]、Liu 等[13]。但这些研究中，都是以等权重方式处理特征指示词和候选特征词之间的关系，没有考虑两者关系的强度，也没有考虑候选词之间的语义相似性对其置信度的影响。

本文在生成候选特征词和候选观点词二分图基础上，综合考虑关联关系和语义关系，利用改进 HITS 算法构建了置信度计算模型，通过置信度排序联合抽取特征词和观点词。

3　基于改进 HITS 的特征观点置信度排序模型

3.1　研究概述

本文研究框架主要任务包括：候选对象提取、二分图构建及关系计算、置信度计算模型构建、实验结果及分析评价等部分，如图 1 所示：

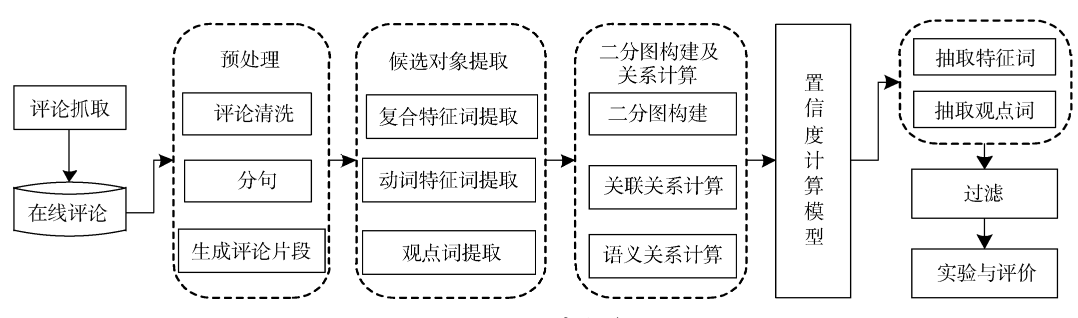


图 1　研究框架

3.2　候选特征观点词提取

相关研究通常选择语料中的名词作为候选特征词，形容词和动词作为候选观点词[11,13]。但通过观察语料会发现，动词也经常作为特征词或复合特征词出现，例如，在手机评论中会出现“通话 v 质量 n 很好，送货 v 也很及时，操作 v 简单”类似评论，如果直接抽取名词作为候选特征词，会将原本的复合特征词拆解为单个特征词和单个观点词，造成特征语义表达不准确，而直接抽取动词作为候选观点词，会造成将特征词当成观点词的抽取错误。因此，综合考虑词语词性和词语依存关系两方面因素，对句子按“先特征、后观点”的分步策略抽取候选特征观点词。

利用依存句法分析器可以同时得到句子中词语的词性及词语间搭配关系，图 2 所示为利用哈尔滨工业大学语言云的句法解析结果。

使用一个三元组 $Triple<A_pos, B_pos, dp>$ 表示词语词性及依存关系对，A_pos 表示词语 A 及其对应词性，dp 表示词语 A 和 B 的依存关系，按以下规

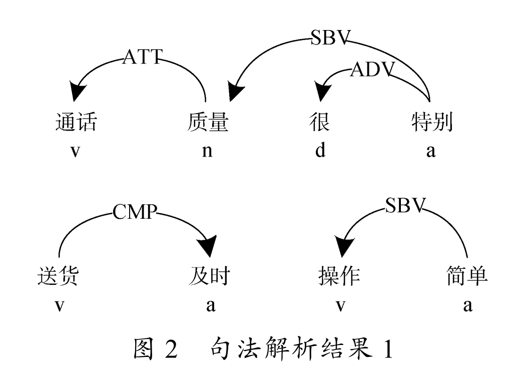


图 2 句法解析结果 1

则进行抽取:

(1) 若 $Triple \in \{<A_n, B_n, ATT>, <A_n, B_v, ATT>, <A_v, B_n, ATT>\}$，则 A 和 B 作为一个复合特征词抽取。例如，对于词对依存关系 $<$外观$_n$, 设计$_n, ATT>$，$<$通话$_v$, 质量$_n, ATT>$，从中抽取复合特征词"外观设计"、"通话质量"。

(2) 若 $Triple \in \{<A_v, B_a, SBV>, <A_v, B_a, CMP>, <A_v, B_a, VOB>\}$，则 A 作为动词特征词抽取，例如，对于词对依存关系 $<$操作$_v$, 简单$_a, SBV>$，$<$显示$_v$, 不错$_a, VOB>$，从中抽取动词特征词"操作"、"显示"。

(3) 对于句子中的其他词语，如果不满足规则(1)和规则(2)，则仅按词性进行抽取，将名词作为候选特征词，形容词和动词作为候选观点词。最后，生成所有候选特征词的集合 T，生成所有观点词的集合 O。

3.3 特征观点二分图构建

在以句子为片段提取了候选特征词和候选观点词后，接下来建立两者的二分图。根据相关文献，可以构建有向二分图[2]，也可以构建无向二分图[13]。考虑到用户在发表评论时是以产品特征为目标对象发表评价观点，观点词是特征词的重要指示词[9]，因而，本文建立一个候选观点词和候选特征词之间的有向二分图。为了便于说明二分图的构建过程，以手机领域的三条评论为例:

①外形小巧，通话质量和做工都不错。

②外形非常小巧轻便，价格也便宜。

③外形特别小巧，非常适合女孩子。另外，价格也很不错。

例如，"外形"、"通话质量"等作为候选特征词抽取出来，"小巧"和"不错"等作为候选观点词抽取出来，每条评论片段内，将所有候选特征词和所有候选观点词连接起来，连接方向为候选观点词指向候选特征词，则建立的二分图如图 3 所示:

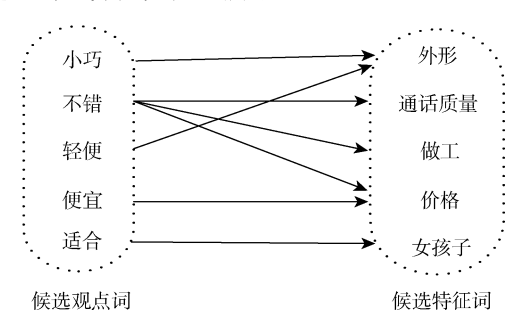


图 3 候选观点词与候选特征词的二分图

3.4 关联关系计算

上节构建的网络图没有体现两个连接节点之间的关联程度的高低。如"外形"与"小巧"共现 3 次，而"外形"和"轻便"共现 1 次，显然前者的关联强度更大，因此本文考查共现的候选特征观点词之间的连接强度。在以"词共现"为基础的关联度计算中，相关研究一般采用互信息法(Mutual Information, MI)度量[14]，因此本文采用互信息值作为候选特征词和候选观点词之间的关联度，设候选特征词为t，候选观点词为o，则两者的关联度计算方法如公式(1)所示。和前文一致，仍采用以评论片段为单位进行关联度计算。

$$I(t,o)=\Pr(t,o)\times\log\frac{\Pr(t,o)}{\Pr(t)\Pr(o)}+\Pr(\neg t,o)\times\log\frac{\Pr(\neg t,o)}{\Pr(\neg t)\Pr(o)}+\Pr(t,\neg o)\times\log\frac{\Pr(t,\neg o)}{\Pr(t)\Pr(\neg o)}+\Pr(\neg t,\neg o)\times\log\frac{\Pr(\neg t,\neg o)}{\Pr(\neg t)\Pr(\neg o)} \tag{1}$$

其中，$I(t,o)$ 表示词 t 和词 o 的关联度，$\Pr(t)$ 和 $\Pr(o)$ 分别表示词 t 和词 o 出现的概率，$\Pr(t,o)$ 表示词 t 和词 o 在语料中联合出现的概率，$\Pr(\neg t,o)$ 和 $\Pr(t,\neg o)$ 表示词 t 和词 o 仅出现其一的联合概率，$\Pr(\neg t,\neg o)$ 表示词 t 和词 o 均未出现的联合概率。

3.5 语义关系计算

语义关系即词语间的语义相似性，借鉴文献[13]，利用对称相对熵(Symmetric Kullback-Leibler)度量词语之间的语义相似性。设有词语 w_i, w_j，两者的语义距离计算公式如下:

$$\begin{aligned}D(w_i,w_j)&=\frac{1}{2}(KL(w_i\|w_j)+KL(w_j\|w_i))\\&=\frac{1}{2}\Big(\sum_{k=1}^{z}p(k|w_i)\log\frac{p(k|w_i)}{p(k|w_j)}+\\&\sum_{k=1}^{z}p(k|w_j)\log\frac{p(k|w_j)}{p(k|w_i)}\Big)\end{aligned} \tag{2}$$

$KL(w_i \| w_j)$ 即 w_i, w_j 的相对熵，表示词 w_i, w_j 在 z 个主题下分布的相异度。其中 $p(k|w_i)$ 通过贝叶斯公式可进一步表示为:

$$p(k|w_i) = p(w_i|k)\frac{p(k)}{p(w_i)} \quad (3)$$

采用 LDA 主题模型估算主题 k 的分布 $p(k)$ 和主题 k 下词 w_i 的分布 $p(w_i|k)$，从而得到 $p(k|w_i)$，同理估算 $p(k|w_j)$。

对于词语为复合词的情形，则分别计算复合词内每个词语与目标对象的每个词语的相对熵，取其最大值进行计算，设 p_i 为复合词，q_j 为目标对象，w_{im} 与 w_{jn} 分别对应 p_i，q_j 内的单词，则复合词语义距离计算公式如下:

$$D(p_i,q_j) = \frac{1}{2}\left(\max_{w_{im}\in p_i, w_{jn}\in q_j}(KL(w_{im}\|w_{jn})) + \max_{w_{jn}\in q_j, w_{im}\in p_i}(KL(w_{jn}\|w_{im}))\right) \quad (4)$$

最后将语义距离进行归一化，得到语义相似性值，用 S 表示 w_i, w_j 的语义相似性，则:

$$S(w_i,w_j) = \frac{1}{1+e^{D(w_i,w_j)}} \quad (5)$$

3.6 置信度排序模型

(1) 考虑关联关系

由于候选特征词与其所关联的候选观点词之间存在双向增强关系，可以应用 HITS 算法迭代计算候选特征词和候选观点词的置信度[2-3]。

为此，在候选特征观点上定义二分有向图 G=(O, T, E), O 表示候选观点词集合, T 表示候选特征词集合, E 表示 O 指向 T 的边集合，用 M 表示图 G 的邻接矩阵，由于本文算法考虑了关联强度，因而需要计算边权重。定义图 G 的关联强度邻接矩阵 M，表示如下:

$$M_{ot} = \begin{cases} I(o,t) & if(o,t)\in E \\ 0 & otherwise \end{cases} \quad (6)$$

其中，矩阵元素 I(o,t)的取值由公式(1)计算得出。

借鉴文献[2]，用集合 T 中节点 t 的 Authority 值表示候选特征词的置信度，记为 A(t)，集合 O 中节点 o 的 Hub 值表示候选观点词的置信度，记为 H(o)，则 A(t) 和 H(o)的计算如下:

$$A(t)= \sum_{(o,t)\in E} H(o) \quad (7)$$

$$H(o)= \sum_{(o,t)\in E} A(t) \quad (8)$$

公式(7)表示节点 t 的置信度由指向 t 的所有节点 o 的当前置信度 H(o)值之和决定，公式(8)表示节点 o 的置信度由 o 指向的所有节点 t 的当前置信度 A(t)值之和决定。

用向量 **A** 表示 T 中所有候选特征节点的置信度，用向量 **H** 表示 O 中所有候选观点节点的置信度，则以矩阵形式描述的置信度计算模型为:

$$\mathbf{A}=M_{ot}^T\mathbf{H} \quad (9)$$

$$\mathbf{H}=M_{ot}\mathbf{A} \quad (10)$$

设 **A** 和 **H** 的初始值为 1 并用 L2 范式规范化处理，通过迭代计算直至算法收敛。

(2) 考虑关联关系和语义关系

语义相似的候选词语之间，其置信度值会相互增强。考虑候选特征词(候选观点词)间的语义相似性作为迭代因子，构建基于关联关系和语义关系的综合置信度计算模型。

利用公式(5)，分别构造基于候选特征词的语义相似度邻接矩阵 M_{tt} 和基于候选观点词的语义相似度邻接矩阵 M_{oo}，其中:

$$M_{tt} = \begin{cases} S(t_i,t_j) & if(t_i,t_j)\in T \text{ and } t_i \neq t_j \\ 0 & otherwise \end{cases} \quad (11)$$

$$M_{oo} = \begin{cases} S(o_i,o_j) & if(o_i,o_j)\in O \text{ and } o_i \neq o_j \\ 0 & otherwise \end{cases} \quad (12)$$

矩阵元素 M_{tt} 和 M_{oo} 的取值由公式(5)计算得到。

构造包含关联关系和语义关系的候选特征观点对置信度计算模型，其矩阵形式表示如下:

$$\mathbf{A}= \lambda M_{ot}^T\mathbf{H}+ (1-\lambda) M_{tt}\mathbf{A} \quad (13)$$

$$\mathbf{H}= \lambda M_{ot}\mathbf{A}+ (1-\lambda) M_{oo}\mathbf{H} \quad (14)$$

模型表示，候选特征词(候选观点词)的置信度由其关联的候选观点词(候选特征词)和其语义相近的候选特征词(候选观点词)的置信度共同决定，其中，λ 为调节参数。迭代运算公式(13)、公式(14)，每次在下一次迭代前，向量 **A**、**H** 值均用 L2 范式进行规范化处理，直至算法收敛。依据候选特征观点词的置信度值排序，设置阈值 γ_t 和 γ_o，分别抽取大于阈值的词语作为特征词集和观点词集。

3.7 特征观点词过滤

实验发现，通过上节得到的特征观点词集合中，还会存在少量泛化名词(如“问题”、“方面”等)和不具有

明显情感极性的动词("打开"、"看到"等), 这是由于一些频繁出现的非特征名词和非观点动词存在较强关联性, 在计算结果中也具有较高置信度, 因而, 有必要将这些置信度高、但却不是抽取对象的词语剔除。为此, 分别对特征词和观点词进行过滤。

(1) 特征词过滤

特征词包括通用特征词和领域相关特征词, 前者指不依赖于特定领域的产品特征, 如"价格"、"服务"等, 而后者指和具体产品相关的特征词, 如手机领域的"听筒"、"蓝牙"等。一般而言, 通用特征词类目和词汇数量较少, 适合于人工构建[15], 为此, 人工定义价格、服务、物流、质量、外观、效果等 6 大类目种子通用特征词, 依据同义词词典词库进行扩展。通过比对特征词集和该词库, 出现在其中的作为通用特征词抽取。

对于领域相关特征词, 其在对应领域中出现的概率, 要比另外产品领域中出现的概率大得多, 因此可以根据特征词在其对应领域和另一不相关领域中出现的概率差值判断[16]。举例来说, 手机评论中频繁出现的"蓝牙"、"通话声音"、"分辨率"等, 不会出现在护肤产品评论集中, 而"习惯"、"情况"、"事情"等泛化词在两个评论集中出现的概率则类似。因此, 对于特征词 t, 计算特征词在两个评论集(D_1, D_2)中出现的概率差值, 将大于阈值 θ 的词语作为领域特征词抽取出来, 如下所示:

$$\mathrm{prob}(t|D_1)-\mathrm{prob}(t|D_2)\geqslant\theta \tag{15}$$

其中, $\mathrm{prob}(t|D_1)$ 、 $\mathrm{prob}(t|D_2)$ 代表词 t 在领域相关评论集 D_1、领域不相关评论集 D_2 中出现的概率。

(2) 观点词过滤

观点词中出现的少量不具有情感极性的词语, 可以运用情感极性过滤方法去除其中不具有明显极性的观点词。相关文献一般采用 HowNet 词库及词语相似性计算方法, 判断观点词的情感极性。然而, HowNet 的词库范围有限, 许多网络新词(如"给力"等)并未包含其中, 对于这类 HowNet 未收录的观点词, 本文采取统计方法判断其情感极性。具体而言, 首先构建大小相同的褒义基准词集 pos_seed 和贬义基准词集 neg_seed, 利用以下公式判断观点词是否具有情感极性:

$$\mathrm{Polarity}(o)=\left|\sum_{i=1}^{|pos_seed|}\mathrm{sim}(w_i,o)-\sum_{j=1}^{|neg_seed|}\mathrm{sim}(w_j,o)\right| \tag{16}$$

其中, $\mathrm{sim}(w_i,o)$ 为观点词与褒义基准词集的语义相似度, $\mathrm{sim}(w_j,o)$ 为观点词与贬义基准词集的语义相似度, 通过 HowNet 计算, 若 Polarity(o) 的绝对值接近于 0, 表示该观点词的情感极性不明显, 若该值显著大于 0, 则该观点词具有明显的情感极性。对于 HowNet 无法判断的观点词, 采用该词与褒义基准词集和贬义基准词集的关联度差值来判断, 公式如下所示:

$$\mathrm{Polarity}(o)=\left|\sum_{i=1}^{|pos_seed|}\log\frac{\mathrm{hits}(o,w_i)}{\mathrm{hit}(o)\mathrm{hit}(w_i)}-\sum_{j=1}^{|neg_seed|}\log\frac{\mathrm{hits}(o,w_j)}{\mathrm{hit}(o)\mathrm{hit}(w_j)}\right| \tag{17}$$

其中, $\mathrm{hits}(o,w_i)$ 表示观点词与褒义基准词的共现频次, hit(o) 与 $\mathrm{hit}(w_i)$ 分别表示观点词与基准词单独出现的频次。若 Polarity(o) 的绝对值接近于 0, 表示该词与褒义词和贬义词关联程度基本相同, 情感极性不明显; 若该值显著大于 0, 则该观点词具有明显的情感极性, 予以保留。

3.8 特征观点对配对与抽取

由于观点词一般修饰其距离最近的特征词, 为此可以考虑将特征词和其最近的观点词进行配对并抽取特征观点对, 考虑在每个评论片段内特征词和观点词可能会出现一对一、一对多、多对一等表达形式, 因此定义 5 种配对模式及抽取规则, 具体如表 1 所示:

表 1 特征观点对抽取模式及示例

序号	评论片段	配对模式	抽取结果
1	屏幕色彩/T…漂亮/O…	TO	(屏幕色彩, 漂亮)
2	…合理/O 的价位/T…	TO	(价位, 合理)
3	…音质/T 和界面/T 都很不错/O…	TO+TO	(音质, 不错), (界面, 不错)
4	…外观/T 漂亮/O…精致/O…	TO+TO	(外观, 漂亮), (外观, 精致)
5	…优雅/O 而小巧 O 的机型 T…	TO+TO	(机型, 优雅), (机型, 小巧)

4 实验与结果分析

4.1 语料来源及预处理

以亚马逊网站评论为实验语料来源, 选择 Nokia 手机和 Canon 相机有效评论(不包括重复和广告评论)作为实验对象, 评论日期截至 2014 年 12 月, 分别选择 1 000 条手机评论和 1 200 条相机评论作为实验语料。

邀请三名具有信息系统研究背景的成员参与标注工作，两名成员对实验语料中的特征词和观点词进行标注。对特征词进行标注时，要求对出现的复合特征词作为一个特征词标注，同时标注出所有特征词的词性，当标注结果不一致时，邀请第三名成员进行校验，随机抽取 50 条语料计算 Kappa 统计量(Cohen, 1960)，以检验标注结果的一致性，结果显示 Kappa 值约为 0.81，表明标注一致性结果可接受。实验语料的统计及标注结果具体如表 2 所示:

表 2 实验语料统计结果

类型	评论总数	评论片段	平均评论长度	特征词	观点词	特征观点对
手机	1 000	2 852	57.3 字/条	278	244	1 764
数码相机	1 200	4 526	46.8 字/条	309	327	2 155

4.2 实验说明

对实验语料划分评论片段，再调用哈尔滨工业大学语言云(LTP-Cloud)的开源 API 接口[17]生成 XML 文件，获取评论片段的分词、词性标注和依存句法分析结果，采用 Python Gensim 包生成 LDA 主题模型，经过实验比较，选择主题 K=12，调节参数取最优值 0.5。使用准确率(P)、召回率(R)和调和平均值(F)对实验结果进行评价。为使算法收敛，得到较为准确的结果，将收敛阈值设为 10^{-5}，即当相邻两次迭代结果之差小于阈值时算法终止。

4.3 实验结果

(1) 特征观点提取结果

按置信度值排序，分别得出手机和数码相机实验评论语料中前 10 个特征词和观点词，如表 3 所示:

表 3 产品特征观点词提取结果

手机		数码相机	
特征词(置信度)	观点词(置信度)	特征词(置信度)	观点词(置信度)
外观(0.097692)	便宜(0.057662)	功能(0.108354)	清晰(0.069875)
屏幕(0.092476)	小巧(0.053501)	像素(0.103563)	不错(0.066367)
质感(0.090291)	实惠(0.050207)	屏幕(0.099891)	简单(0.063258)
功能(0.088795)	精致(0.048622)	镜头(0.098679)	漂亮(0.059324)
手写功能(0.086416)	方便(0.046619)	画质(0.097653)	喜欢(0.057921)
性价比(0.085824)	漂亮(0.042588)	相片(0.092835)	容易(0.055346)
手感(0.085312)	高(0.040548)	色彩(0.090178)	好(0.051789)
价格(0.083047)	实用(0.039778)	价格(0.088546)	满意(0.050328)
品牌(0.081091)	简单(0.038319)	单反(0.087193)	一般(0.049765)
款式(0.079978)	满意(0.038088)	效果(0.085649)	清楚(0.048561)

(2) 按置信度模型抽取结果统计

比较不同阈值下实验数据的特征观点词的识别精度，最终确定候选特征词和候选观点词的置信度阈值，其结果分别如表 4 和表 5 所示:

表 4 特征词抽取结果

类型	置信度阈值	抽取特征数	准确数	准确率	召回率
手机	0.035	284	237	0.835	0.853
相机	0.033	313	249	0.796	0.806

表 5 观点词抽取结果

类型	置信度阈值	抽取观点数	准确数	准确率	召回率
手机	0.014	264	202	0.765	0.828
相机	0.015	335	241	0.719	0.737

4.4 对比实验

为了验证本文方法的有效性，选择 Aravindan 等[10]和 Zhang 等[2]两个代表性的研究方法(分别称为方法1和方法2)，和本文方法(称为方法3)进行对比实验；另一方面，基于本文方法过滤策略设计方法4，验证特征观点过滤策略的有效性。采用准确率、召回率、调和平均值作为评价指标。

(1) 方法1

抽取实验语料所有名词对象作为候选特征词，采用 Apriori 算法找出 1 项频繁特征集和 2 项频繁特征集，由于中文评论中较少出现 3 项及以上频繁特征集，因此不考虑 3 项及以上频繁特征集。参照文献[10]，设置项集最小支持度为 0.01，置信度为 0.8。采用近邻规则对 2 项频繁特征集进行过滤，采用独立支持度对 1 项频繁特征集进行过滤。过滤后，得到所有特征词，抽取其最近的形容词或动词作为观点词，并按 3.8 节定义模式抽取特征观点对。

(2) 方法2

采用文献[2]中提出的方法，利用 HITS 算法排序抽取特征词。文献[2]中没有抽取观点词，因此抽取特征词最近的形容词或动词作为观点词，并按 3.8 节定义模式抽取特征观点对。

(3) 方法3

基于本文置信度模型，设置阈值抽取特征观点集合。

(4) 方法4

在方法 3 基础上，应用 3.7 节过滤策略，进行特征

观点的再过滤。分别在手机和数码相机实验语料上进行对比实验，结果如表 6 和表 7 所示:

表 6 手机数据语料实验结果

方法	特征词			观点词			特征观点对		
	P	R	F	P	R	F	P	R	F
1	0.727	0.736	0.731	0.704	0.714	0.709	0.672	0.668	0.670
2	0.756	0.817	0.785	0.712	0.735	0.723	0.691	0.673	0.682
3	0.835	**0.853**	0.845	0.765	**0.828**	0.795	0.734	0.756	0.745
4	**0.857**	0.845	**0.851**	**0.810**	0.824	**0.817**	**0.753**	**0.769**	**0.761**

表 7 数码相机语料实验结果

方法	特征词			观点词			特征观点对		
	P	R	F	P	R	F	P	R	F
1	0.693	0.702	0.697	0.668	0.683	0.675	0.615	0.652	0.633
2	0.682	0.713	0.697	0.653	0.662	0.657	0.601	0.629	0.615
3	0.796	**0.806**	0.801	0.719	**0.737**	0.728	0.695	0.701	0.698
4	**0.825**	0.796	**0.810**	**0.756**	0.728	**0.742**	**0.729**	**0.717**	**0.723**

通过表 6 和表 7 的实验对比结果，可以看出:

(1) 经过两组语料的实验分析，在特征词和观点词的抽取效果上，方法3都优于两组基线方法，说明了本文方法在特征词和观点词识别上的有效性。

(2) 基线方法1的准确率和召回率都较低，说明使用频繁特征词方法并不能有效抽取所有特征词，主要原因是方法1采用名词及名词短语作为候选特征词，没有考虑语料中的动词特征词的抽取，从而影响了特征词和观点词的抽取准确率和召回率。

(3) 基线方法2的实验结果略高于方法1，说明利用 HITS 算法提取特征词的方法具有有效性。和方法3比较，方法2中没有考虑关系强度和语义关系等因素，实验效果低于本文方法，说明关系强度和语义关系对于识别候选对象具有一定效果。

(4) 采用过滤策略的方法4在两组实验语料上均取得较高的准确率，说明特征词和观点词的过滤策略具有一定的有效性。比较而言，观点词的准确率有较大提升，反映出利用情感极性进行观点词过滤作用明显。同时，和实验3相比，两组语料的召回率略有下降，但总体来看，采用过滤后的特征观点词提取特征观点对，能取得更好的实验准确率和召回率。

5 结 语

面对海量的在线评论，如何克服其口语化严重、表达不规范的特点，有效识别出产品特征词和观点词具有重要的应用价值，可以应用于电子商务、舆情监控、客户知识管理、竞争情报分析等领域。本文基于相互增强关系的思想，利用改进 HITS 算法构建置信度排序模型抽取中文评论中的特征词和观点词。首先考虑动词特征词抽取策略，避免动词特征词的遗漏，以及特征词识别召回率不高的问题，在置信度计算模型中，本文不仅考虑候选特征词和候选观点之间的共现关系，还考虑候选特征词之间、候选观点词之间的语义关系。以手机语料为分析对象的实验结果表明，综合关联关系和语义关系的分析框架，利用置信度排序模型抽取特征词和观点词具有较高的准确率，具有一定的有效性。

本文主要考虑的是显性特征词和观点词的识别，然而，在线评论中还包含一定数量的隐性特征词，由于篇幅原因，并未对隐性特征词的提取进行讨论，后续研究将针对这一问题展开。

(致谢：本文研究中使用了哈尔滨工业大学和科大讯飞股份有限公司的“哈工大–讯飞语言云”接口，在此表示感谢!)

参考文献:

[1] 王永，张勤，杨晓洁. 中文网络评论中产品特征提取方法研究[J]. 现代图书情报技术，2013(12): 70-73. (Wang Yong, Zhang Qin, Yang Xiaojie. Research on the Method of Extracting Features from Chinese Product Reviews on the Internet [J]. New Technology of Library and Information Service, 2013(12): 70-73.)

[2] Zhang L, Liu B, Lim S H, et al. Extracting and Ranking Product Features in Opinion Documents [C]. In: Proceedings of the 23rd International Conference on Computational Lingusitics (COLING), Beijing, China. Stroudsburg, PA, USA: ACL, 2010: 1462-1470.

[3] 郗亚辉. 产品评论特征及观点抽取研究[J]. 情报学报，2014, 33(3): 326-336. (Xi Yahui. Extracting Product Features and Opinions from Product Reviews [J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(3): 326-336.)

[4] Jin W, Ho H H, Srihari R K. A Novel Lexicalized HMM-based Learning Framework for Web Opinion Mining [C]. In: Proceedings of the 26th Annual International Conference on Machine Learning (ICML), Montreal, Canada. New York, NY, USA: ACM, 2009: 465-472.

[5] Li F T, Han C, Huang M L, et al. Structure-aware Review Mining and Summarization [C]. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING), Beijing, China. Stroudsburg, PA, USA: ACL, 2010: 653-661.

[6] Wu Y B, Zhang Q, Huang X J, et al. Phrase Dependency Parsing for Opinion Mining [C]. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP), Singapore. Morristown, NJ, USA: ACL, 2009: 1533-1541.

[7] Titov I, McDonald R. Modeling Online Reviews with Multi-grain Topic Models [C]. In: Proceedings of the 17th International Conference on World Wide Web (WWW), Beijing, China. New York, NY, USA: ACM, 2008: 111-120.

[8] Zhao W X, Jiang J, Yan H F, et al. Jointly Modeling Aspects and Opinions with a MaxEnt-LDA Hybrid [C]. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP), Massachusetts, USA. Stroudsburg, PA, USA: ACL, 2010: 56-65.

[9] Hu M Q, Liu B. Mining and Summarizing Customer Reviews[C]. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Seattle, USA. New York, NY, USA: ACM, 2004: 168-177.

[10] Aravindan S, Ekbal A. Feature Extraction and Opinion Mining in Online Product Reviews [C]. In: Proceedings of the 2014 International Conference on Information Technology (ICIT), Bhubaneswar, India. New York, NY, USA: IEEE, 2014: 94-99.

[11] Qiu G, Liu B, Bu J J, et al. Opinion Word Expansion and Target Extraction Through Double Propagation [J]. Computational Linguistics, 2011, 37(1): 9-27.

[12] Hai Z, Chang K Y, Cong G. An Association-Based Unified Framework for Mining Features and Opinion Words [J]. ACM Transaction on Intelligent Systems and Technology, 2015, 6(2): 2601-2626.

[13] Liu K, Xu L H, Zhao J. Extracting Opinion Targets and Opinon Words from Online Reviews with Graph Co-ranking [C]. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Ligustics (ACL), Baltimore, USA. Stroudsburg, PA, USA: ACL, 2014: 314-324.

[14] 尹裴，王洪伟，郭恺强．中文产品评论的“特征观点对”识别：基于领域本体的建模方法[J]．系统工程，2013，31(1): 68-77. (Yin Pei, Wang Hongwei, Guo Kaiqiang. Feature-opinion Pair Identification in Chinese Online Reviews Based on Domain Ontology Modeling Method [J]. Systems Engineering, 2013, 31(1): 68-77.)

[15] 郑波，胡其，林君．文本语义分析的实现及应用[J]．程序员，2013(7): 105-109. (Zheng Bo, Hu Qi, Lin Jun. Implementation and Application of Semantic Analysis in Text [J]. Programmer, 2013(7): 105-109.)

[16] Kansal H, Toshniwal D. Aspect Based Summarization of Context Dependent Opinion Words [J]. Procedia Computer Science, 2014, 35: 166-175.

[17] 哈工大–讯飞语言云. Web Service 接口[EB/OL]. [2015-05-01]. http://ltpapi.voicecloud.cn/. (LTP-Cloud RESTful API [EB/OL]. [2015-05-01]. http://ltpapi.voicecloud.cn/.)

作者贡献声明:

孟园：采集、清洗和分析数据，论文起草及最终版本修订；
王洪伟：论题拟定，提出研究思路，设计研究方案，修改论文。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储，可通过电子邮件向作者索取，E-mail: nancymeng5544@163.com。

[1] 孟园，王洪伟. confidence_ranking_algorithm.py. 改进置信度排序模型算法.

[2] 孟园，王洪伟. experimental_data.xls. 原始实验数据.

[3] 孟园，王洪伟. experimental_data.pkl. 实验数据预处理后生成二级列表数据.

[4] 孟园，王洪伟. gen_adjacency.py. 生成候选词关联关系和语义关系矩阵算法.

[5] 孟园，王洪伟. MM.npy. 候选特征观点词间关联关系矩阵.

[6] 孟园，王洪伟. Mhh.npy. 候选观点词语义关系矩阵.

[7] 孟园，王洪伟. Maa.npy. 候选特征词语义关系矩阵.

[8] 孟园，王洪伟. gen_xml.py. 调用 LTP_Cloud 接口生成 XML 文件程序.

[9] 孟园，王洪伟. get_candidate_from_xml.py. 从 XML 文件提取候选特征观点词程序.

收稿日期: 2015-08-28
收修改稿日期: 2015-10-09

Extracting Product Feature and User Opinion from Chinese Reviews

Meng Yuan　Wang Hongwei
(School of Economics and Management, Tongji University, Shanghai 210000, China)

Abstract: [**Objective**] This study proposed a confidence ranking model to extract product feature and user opinion from the Chinese online reviews. [**Methods**] Examining the semantic and association relations between candidate words, we built the confidence ranking model based on the improved HITS algorithm, and then retrieved the feature and opinion words. [**Results**] Compared with the reference model, our method showed better recall and precision rates while extracting the feature and opinion words from the Chinese corpus. [**Limitations**] Only extracted the explicit feature and opinion words, and did not try to identify and extract the implicit ones. [**Conclusions**] We could effectively extract the feature and opinion words using their mutual reinforcement and semantic relations. Filtering method of the semantic polarity could also improve the precision of the extracted opinion words.

Keywords: Confidence ranking　HITS　Association relation　Semantic relation　Mutual reinforcement　Feature opinion extraction

Wiley 与 Figshare 合作促进数据共享

John Wiley国际出版公司2015年6月宣布与位于伦敦的数据存储库组织Figshare建立合作伙伴关系。为支持有意公开分享其数据的作者, Wiley已经着手与合作伙伴Figshare对现有的期刊工作流程和文章出版物进行面向数据共享的整合。新的数据共享服务将在一批期刊中进行试点, 并在未来的几个月中伴随新的数据引用和数据共享政策逐步推开。这将确保作者和读者可以在知识共享许可协议下免费访问、共享和复制来自Wiley在线图书馆文章中的更多数据。

随着学术研究资助者对数据开放和可获取性要求的不断增长, 提供合乎规范的优化工作流程服务变得越来越重要。这种伙伴关系使得Wiley作为学术内容传播专家在增加研究曝光度的同时仍能继续为作者提供全面综合的发布服务, 也使得Figshare能够提供更强大的数据存储和引用服务。

Wiley之前做过一项关于研究者的需求随着研究进程和新技术的发展而不断变化的广泛调研。随着研究和技术的融合, 让数据可以被人类和机器同时阅读的需求已经成为一个重要的新兴领域。数据的这种灵活性将使得学术研究人员能够更加方便地使用它们。

Figshare的首席执行官Mark Hahnel说: “在以前瞻性思维对数据进行重要投入之前, Wiley对其进行了深入研究。这一做法说明了学术界不断变化的特性以及Wiley对其作者提供世界一流的服务的承诺。由于学术信用体系的发展, 我们想要保证所有的学者都能够得到与其所做工作相对应的声望。这种合作伙伴关系意味着在Wiley发表学术成果的作者将享受到全面的益处。”

Wiley 负责期刊编辑发展的副总裁 Liz Ferguson 指出: “我们一直在寻找为我们的作者提供最具创新性的和最有益的出版体验。资助者为作者增加了许多必须遵守的新的要求。我们的作者服务是无与伦比的, 并且在与 Figshare 合作后将进一步为在 Wiley 期刊中发布成果的学术研究人员提升服务水平。”

(编译自: http://www.wiley.com/WileyCDA/PressRelease/pressReleaseId-119082.html)

(本刊讯)